Automatic and interactive rule inference without ground truth

Cérès Carton, Aurélie Lemaitre, Bertrand Coüasnon

Automatic and interactive rule inference without ground truth

Cérès Carton
IRISA - INSA
Université Européenne de Bretagne
Campus de Beaulieu
35042 Rennes Cedex, France
ceres.carton@irisa.fr

Aurélie Lemaitre
IRISA - Université Rennes 2
Université Européenne de Bretagne
Campus de Beaulieu
35042 Rennes Cedex, France
aurelie.lemaitre@irisa.fr

Bertrand Coüasnon
IRISA - INSA
Université Européenne de Bretagne
Campus de Beaulieu
35042 Rennes Cedex, France
couasnon@irisa.fr

***Abstract*—Dealing with non annotated documents for the design of a document recognition system is not an easy task. In general, statistical methods cannot learn without an annotated ground truth, unlike syntactical methods. However their ability to deal with non annotated data comes from the fact that the description is manually made by a user. The adaptation to a new kind of document is then tedious as the whole manual process of extraction of knowledge has to be redone. In this paper, we propose a method to extract knowledge and generate rules without any ground truth. Using large volume of non annotated documents, it is possible to study redundancies of some extracted elements in the document images. The redundancy is exploited through an automatic clustering algorithm. An interaction with the user brings semantic to the detected clusters. In this work, the extracted elements are some keywords extracted with word spotting. This approach has been applied to old marriage record field detection on the FamilySearch HIP2013 competition database. The results demonstrate that we successfully automatically infer rules from non annotated documents using the redundancy of extracted elements of the documents.**

I. Introduction

Learning automatically document structure recognition system and dealing with the lack of annotated learning database are two difficult tasks themselves. From our knowledge, there is currently no existing method that can tackle both tasks jointly. Learning automatically document structure recognition system is in the general the privileged field of the statistical methods. This learning step makes them easily adaptable to the recognition of a new kind of documents. However, these methods need an annotated ground truth for the learning step. It is the case of methods based on Conditional Random Field [1], 2D Markovian Random Field [2] or EM-based method [3]. Unfortunately, data ground truth is not always available to train the recognition systems. Indeed, data ground truth generation is a laborious and expensive task as it implies human intervention. Methods have been proposed to synthesize ground truth by degradation [4], but they are not adapted for document structure recognition and do not produce realistic documents for this task. No answer is currently given to overcome the need of ground truth for statistical methods.

On the opposite, syntactical methods are in general able to deal with non annotated database. However, this ability comes from the fact that the knowledge on documents is not learned automatically but manually and explicitly expressed as rules by the user [5]. Thus, these methods cannot be easily adapted to a new kind of documents as the whole manual extraction of knowledge has to be redone. Rule inference methods exist and have been studied in numerous fields [6], however it is a very hard task for two dimensional grammatical descriptions. Shilman [7] presents a method to learn non-generative grammatical models for document analysis. They focus their effort on feature selection and parameter estimation. However, this method needs an annotated ground truth to set all the parameters of the model.

In document analysis field, document classification methods have been developed that do not need annotated ground truth or very few documents annotated. In the trend of reducing the amount of annotated data, Rusiñol [8] proposed an incremental method that only needs the manual annotation of one image per provider. In this method, the input documents are OCRed. This method works well if the documents are not too degraded, in order for the OCR to obtain good results, and if we already know all the classes of documents we want to classify. However, document classification methods have not been designed to generate rules to describe the structure of the documents which is our final objective.

In this paper, we propose a method to extract knowledge and infer rules from non annotated documents. The analysis is based on the study of redundancies of extracted elements from documents in large databases. The main advantage of such a method is to cope with the lack of ground truth while providing an automatic and interactive extraction of knowledge. In section II, we present an overview of the proposed method. Then, we present the rule generation from non annotated documents in section III. Finally, we will present in section IV our experiments on the FamilySearch HIP2013 competition database and demonstrate that we successfully generate rules from non annotated documents using the redundancy of the grammar terminals extracted from the documents.

II. Method overview

In order to be able to automatically extract knowledge from non annotated documents, we propose to study redundancies of some extracted elements in document image in a large amount of data. The extracted elements are produced as a previous stage of the analysis, using for example a first general recognition system that does not rely on the knowledge of document structure. These elements can be various: OCR

results, keywords detected with word spotting, text lines or text blocks, line segments, etc. They correspond to the grammar terminals of the rules we want to infer.

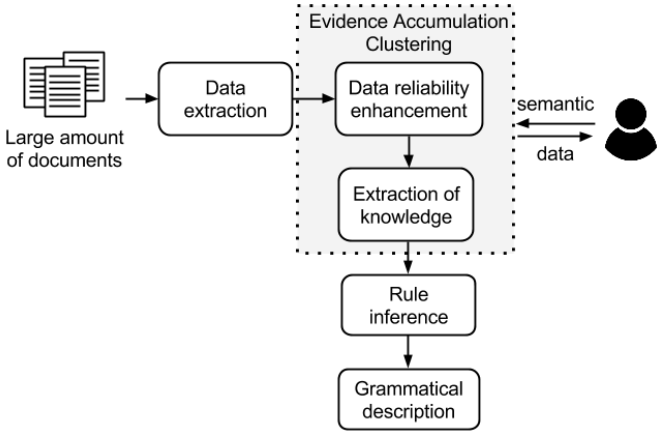

Fig. 1. Rule inference without ground truth: method overview

One of the major properties of these extracted elements is that they are not completely reliable. Amongst all the extracted elements some of them tally with what the user has expected while some others do not. As a result the extraction of knowledge must be preceded by a *data reliability enhancement* step. The reliable elements can then be used for the extraction of knowledge and the rule inference. Both data reliability enhancement and extraction of knowledge are based on the use of clustering techniques and interaction with the user that will brings sense to the data. We detail these steps in the rest of this article.

III. Rule inference from non annotated document

To illustrate our method, we use the FamilySearch HIP2013 competition database of handwritten Mexican marriage records. This database is composed of 10,500 training images and 20,000 test images. The task of the competition can be decomposed in two steps:

1) Localize 4 fields in each record: two fields in a text paragraph (month and year) and two fields in a tabular structure (hometown of groom and bride)
2) Group fields according to their content

The ground truth contains the transcription of fields content for each document but *no localization* of these fields is available. As an illustration of the method, we focus our interest on a sub-task of the competition: the *localization* of the two fields "month" and "year" which are in the text paragraph (figure 6). We thus do not have an available ground truth for our task.

In this example, the extracted elements are 8 keywords selected in the printed text that delimit the handwritten fields we try to locate. They are extracted using the word spotting presented in [9]. An OCR has been tested on the document but has not been selected. Indeed, the documents are too degraded and there are too much interaction between handwritten text and typed text to allow a good recognition. A perfect keywords extraction would lead to obtain one occurrence of each keyword per document, as it can be seen in the example of figure 2. However, as it can be observed in table I, much more occurrences are obtained. This confirms that extracted elements, contrary to the ground truth, are *not directly reliable*. To extract knowledge from these elements, we need to first improve their reliability by removing the occurrences that introduce noise in our analysis.

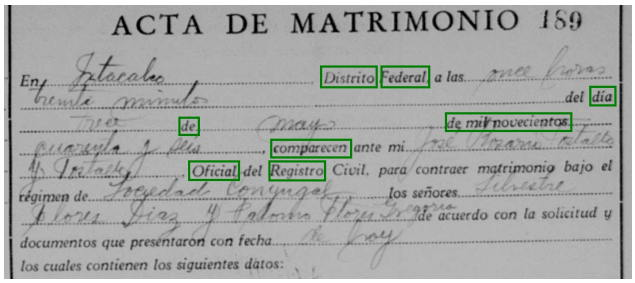

Fig. 2. Representation of the bounding boxes of the eight keywords searched in the documents

Keyword	Number of occurrences
compar	4.0
de	9.1
de mil novecientos	1.1
dia	6.5
Distrito	2.7
Federal	2.0
Oficial	3.7
Registro	1.0

TABLE I. Mean number of keywords occurrences per document

To do so, we use a clustering algorithm: the clustering foregrounds clusters of similar elements and the user then gives meaning to each cluster. Several constraints exist that determined our clustering algorithm choice. First, the clustering algorithm needs to be able to adapt to many different kinds of data sets as our method is generic. Then, we have no a priori on the data, so the clustering algorithm need to have no or few parameters to fix and especially the number of clusters. Finally, we use the clustering at two steps of the analysis: the reliability enhancement and the knowledge extraction (figure 1). To do so, it is of interest to automatically determine an optimal number of clusters for the data partition but also to be able to easily over-segment the data, especially for the data reliability enhancement.

A. Evidence Accumulation Clustering

Considering these requirements, we selected the Evidence Clustering Algorithm introduced by Fred and Jain [10]. The Evidence Clustering Algorithm *builds N partitions* with different clustering algorithms. In our implementation we use one algorithm, the k-means algorithm, with random k for the initialization. The different partitions are combined to *generate a similarity matrix* M between the occurrences:

$$M_{(i,j)} = \frac{n_{ij}}{N} \quad \text{where } n_{ij} \text{ is the number of times } i \text{ and } j \text{ are in the same cluster}$$

The final data partition is obtained through a *hierarchical agglomerative clustering* where *maximum cluster lifetime criterion* is used to cut the dendrogram and then determine the number of clusters of the final partition. It is also possible to easily over-segment the data by choosing an other cutting point in the dendrogram.

B. Data reliability enhancement

As we presented above, the extracted elements are not directly reliable. To improve their reliability, we propose to use the Evidence Clustering Algorithm presented in section III-A.

1) Clustering: For the data reliability enhancement, the clustering is made on the raw extracted elements. The clustering algorithm produces a data partition of the extracted elements that can be used to efficiently and quickly remove whole clusters of occurrences which are not of interest for the extraction of knowledge.

2) Interaction: When the data partition has been built, each cluster has to be visualized by the user. It is an essential step as the user is able to bring sense to the data. He visualizes and decides if they should be kept for the further analyses. To do so efficiently, five representative examples of the cluster are selected and presented to the user. We consider as representative examples of the cluster the occurrences which are close to the centroid of the cluster. In addition to the examples, a measure of the intra-cluster variability is displayed to the user to insure that the examples presented to the user are really representative of the whole cluster occurrences. The smaller the intra-cluster variability, the more confident the user can be in his decision. The user can also visualize more examples if he is not sure of his decision, if the intra-variability is not small for example. The results are then presented from the closest to the centroid to the furthest.

On the example of the Mexican marriage records, the reliability enhancement is performed for each of the eight keywords separately. The clustering is performed using the position in the page on the X and Y axis. The clustering leads to the constitution of partitions of ten to twenty clusters per keyword.

At the end of the reliability enhancement step, the data can be used for the knowledge extraction step. The reliable extracted elements are then used like an *approximate ground truth* of the documents. They correspond to the grammar terminals of the rules we want to infer.

C. Extraction of knowledge

The extraction of knowledge step consists in the study of the redundancies in the reliable extracted elements. The study of the redundancies will allow the detection of structure in our data that can be used for the rule inference and then for the building of an appropriate document recognition system.

The extraction of knowledge can concern any properties that we have on the data. We can study the position of elements (in the page or relatively to each other), the size of the elements, the vocabulary used, etc. This study will give us knowledge on both the logical and the physical structure of the documents to analyze.

In the example on the handwritten Mexican marriage records, we search the position of keywords that delimit the text fields we search to recognize. We first study the position of the eight keywords in the page and then study the global document model.

1) Keyword level analysis: (1) An automatic clustering of the reliable occurrences is made using as variables the position in the page on the X and Y axis. This clustering is performed on the *reliable data* obtained from section III-B. The clustering allows us to detect the model(s) of position per keyword. (2) The results of the clustering are visualized by the user to gives meaning to each cluster. He validates the pertinence of the automatic data clustering. The interaction process is identical to the process describes in section III-B2. The numbers of position models for each keyword are presented in table II. For the "dia" keyword, seven different models are detected. Figure 3 presents examples for two of them.

Keyword	Nb of position models
compar	9
de	5
de mil novecientos	6
dia	7
Distrito	5
Federal	4
Oficial	7
Registro	4

TABLE II. NUMBER OF POSITION MODELS DETECTED FOR EACH KEYWORDS DURING EXTRACTION OF KNOWLEDGE

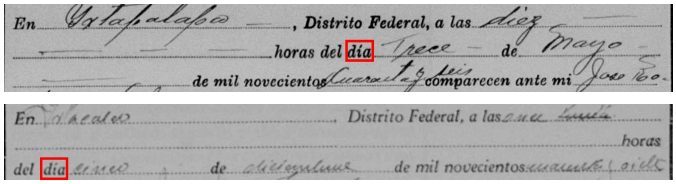

Fig. 3. Examples of two different models of position for keyword "dia" in the FamilySearch competition data set

2) Document level analysis: For each document, a signature is built. To do so, we use the position models detected for each keyword as presented in table III. The keyword occurrence contained in the document is affected to its corresponding position model and the signature is the concatenation of each of the affected position models.

filename	Affected position model dia	de	de mil nov	compar	signature
00001	3	4	7	6	3#4#7#6

TABLE III. EXAMPLE OF CREATION OF THE SIGNATURE FOR A DOCUMENT

To create these signatures, we only use documents where there is one and only one occurrence for each keyword, i.e. documents where there is no ambiguity on the affected position models. The analysis has been performed on 5406 documents on the 7000 documents of the learning data set. Then using these signatures, a frequency analysis of the signature is made. Doing so, 11 models of documents are detected. Figure 4 presents two models of documents. Table IV shows the unequal distribution of the document models in the 5406 documents used of the learning data set.

Model	1	2	3	4	5	6	7	8	9	10	11
Count	1448	822	740	652	566	470	359	123	92	33	25

TABLE IV. THE DOCUMENTS MODELS ARE UNEQUALLY DISTRIBUTED IN THE DATABASE

This unequal distribution of the document models shows the *interest of an automatic and exhaustive analysis* of the database. It would have been very difficult and laborious with a manual analysis of the database to detect all these models. Indeed, when a manual extraction of knowledge is done, only a small sample of documents is analyzed. The chances to obtain a representative small sample are very weak with an unequal distribution like the one we observed on this database.

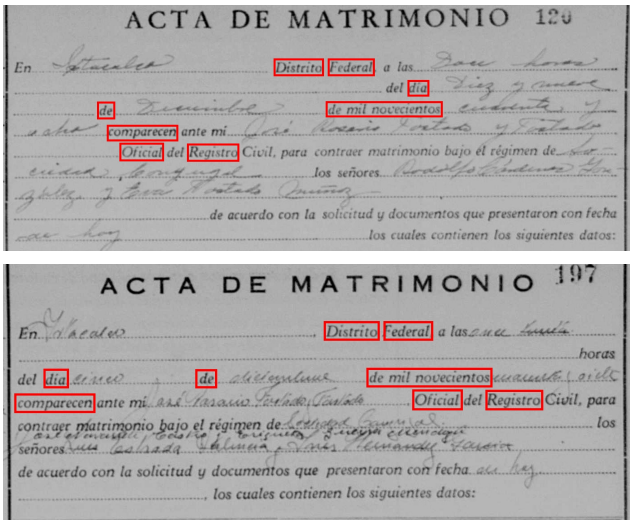

Fig. 4. Examples of two different models of documents in the FamilySearch HIP2013 competition data set

D. Rule inference

After the extraction of knowledge, which has been validated in interaction with the user, we can automatically generate the rules that will be used in the grammatical description. The information that can be inferred is various, concerning both the logical and the physical structure of the documents. The rule inference is made using the *approximate ground truth* that was produced thanks to the data reliability enhancement.

In our example, we need to generate the rules that will describe each document model. Describing one model consists in describing the position of the eight keywords for this specific model. To do so, we need to define 8 positions operators, one for each different keyword. The position operators are inferred following the method presented in [11] with ground truth. It consists in the inference of the 6 parameters that compose them (definition of the zone boundaries and definition of the order of analysis of the elements in the zone). We then automatically infer: 6 parameters $\times$ 8 position operators $\times$ 11 document models = 528 parameters. The generated rules allow building the zones containing the handwritten fields to recognize.

IV. Experimental results

Two different aspects have been evaluated with the experiments. First, we assess our method by evaluating the results obtained with the grammatical description built using the rules inferred without ground truth. Then, we compared the results of our method with the results obtained by the document recognition system submitted for the FamilySearch HIP2013 competition by Lemaitre [9], where four models where manually described.

A. Introduction of the detected models in a grammatical description

The inferred rule can now be integrated in any grammatical system. To validate our method, we used the DMOS method [5] which is a grammatical method for structured document recognition. When all the models have been described, we need to be able to determine which model best fits to the document we are analyzing. To do so, we introduce in the grammatical description the FIND_BEST_FIRST operator presented by Maroneze [12]. When we analyze a document, each of the 11 models of pre-printed forms is tested over the document. A penalty is then computed representing the non-matching to the model. The FIND_BEST_FIRST operator allows us to select the model that best fits the document, which is the model with the lowest penalty.

Each keyword is searched at its supposed position in the model. If the keyword is not found, then the penalty for the model is increased by one. If the keyword is found, then the penalty for the model is increased by:

$$1 - \frac{\textit{intersection area}}{\textit{keyword area}}$$

Figure 5 shows an example of penalty computation. The “de mil novecientos” keyword is searched in the red area. An occurrence is found which is not totally included in the research zone. The computed penalty is 0.477444.

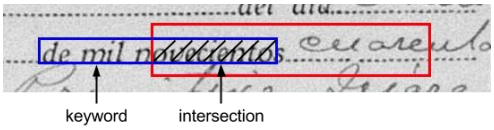

Fig. 5. The ”de mil novecientos” keyword penalty is 0.477444 as it is not totally inside the search zone

B. Database

To evaluate the results obtained by our method, we have annotated 2000 documents of the FamilySearch competition test data set. What is annotated in our document is the position of the “month” and “year” fields in the document and not the position of the text that corresponds to these fields in the documents as it is shown in figure 6. Even if no text is contained in the bottom left “year” zone, it is annotated as we are not searching the text but field position in the page.

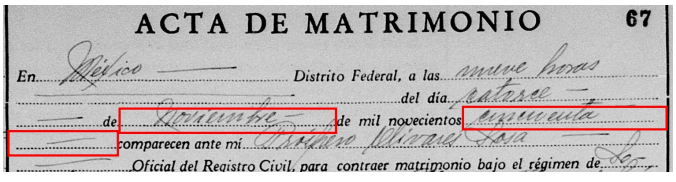

Fig. 6. Example of ground truth annotation for a document: the “month” field is composed of one zone and the “year” field is composed of two zones

C. Metric

The FamilySearch HIP2013 competition could not be used as we focus our work on a sub-task of the competition: the localization of the “month” and “year” fields of each record.

To evaluate this sub-task, we need to evaluate the spatial correspondence of the zones produced by our method with the ground truth zones. To do so, we use the metric introduced by Garris [13]. We want to evaluate the coverage i.e. how well the hypothesis zones cover the ground truth zones. The width of the intersection zones must be very close to the width of the references zones and the intersection height must also be sufficiently big.

We define two sets of thresholds to evaluate our results:

- a field is *completely recognized* if at least 95% of its width and 75% of its height have been recognized
- a field is *partially recognized* if 1) it is not totally recognized, 2) at least 80% of its width has been recognized and 75% of its height.

In other cases, the field will be reported as missing.

D. Results

To show the interest of our method to infer rules without ground truth, we first compare the fields obtained with our method with the ground truth fields. The results presented in table V show that our method efficiently localizes the two searched fields in the documents: only 2.4% of the fields are not detected and 89.8% of the documents are correctly recognized (i.e. all zones are completely or partially recognized in the document). This demonstrates that the grammatical description automatically inferred without ground truth is efficient and accurate.

		Our method	Manual models
Zone	Complete recognition	91.4%	89.7%
	Partial recognition	6.2%	4.0%
	Missing	**2.4%**	6.3%
Document recognition rate		**89.8%**	78.9%

TABLE V. COMPARISON OF THE RESULTS OBTAINED WITH OUR METHOD AND THE METHOD BASED ON 4 MANUALLY DEFINED MODELS

When we compare our results with the results obtained with the method based on manually defined models, we can observe that our method is more accurate and that there are less missing zones (131 missing zones compare to 343 zones with the method based on manually defined models). As we presented it in section III, the detected keywords are not reliable. They can be absent or detected at a wrong place. With the competition method, the document models are not sufficiently precise to overcome the unreliability of the keywords and select the correct ones for the document. For example, model C described by Lemaitre [9] corresponds to six different models in our method. As a result, our method is able to correctly recognized 11% documents more than the method based on manually defined models (217 documents).

V. CONCLUSION

In this paper we have presented a method to infer rules from non annotated documents. To overcome the lack of annotated ground truth, we based our analysis on the study of redundancies of extracted elements in large data set. This analysis combines a reliability enhancement of the extracted elements with an automatic extraction of knowledge. For both of these tasks, we successfully used the Evidence Accumulation Clustering presented by Fred and Jain [10]. The interaction with the user allows bringing sense to the data put forward by the clustering algorithm.

We have validated our work on the FamilySearch HIP2013 competition database, focusing our work on a sub-task of the competition, the localization of "month" and "year" fields. 2,000 documents of the test database have been manually annotated for this purpose that will be publicly available. Our result show that we can efficiently extract knowledge from non annotated documents which is a great improvement as manual annotation of documents is very costly step in the design of document recognition systems. The introduction of an iterative mechanism could improve this mechanism by automatically detecting interested occurrences that would have been deleted during the data reliability enhancement step. This iterative step could also be used to detect new configuration of documents to improve the document recognition system.

REFERENCES

[1] F. Montreuil, S. Nicolas, E. Grosicki, and L. Heutte, "A new hierarchical handwritten document layout extraction based on conditional random field modeling," in *Frontiers in Handwriting Recognition (ICFHR), 2010 International Conference on*, Nov 2010, pp. 31–36.

[2] M. Lemaitre, E. Grosicki, E. Geoffrois, and F. Preteux, "Preliminary experiments in layout analysis of handwritten letters based on textural and spatial information and a 2d markovian approach," in *Document Analysis and Recognition, 2007. ICDAR 2007. Ninth International Conference on*, vol. 2, Sept 2007, pp. 1023–1027.

[3] F. Cruz and O. R. Terrades, "Em-based layout analysis method for structured documents," in *22nd International Conference on Pattern Recognition*, 2014, pp. 315–320.

[4] V. C. Kieu, N. Journet, M. Visani, J.-P. Domenger, and R. Mullot, "Semi-synthetic Document Image Generation Using Texture Mapping on Scanned 3D Document Shapes," in *International Conference on Document Analysis and Recognition (ICDAR 2013)*, Washington, DC, United States, Aug. 2013, pp. 489–493.

[5] B. Coüasnon, "Dmos, a generic document recognition method: Application to table structure analysis in a general and in a specific way," *International Journal on Document Analysis and Recognition, IJDAR*, vol. 8, no. 2, pp. 111–122, June 2006.

[6] C. de la Higuera, "A bibliographical study of grammatical inference," *Pattern Recogn.*, vol. 38, no. 9, pp. 1332–1348, Sep. 2005.

[7] M. Shilman, P. Liang, and P. Viola, "Learning nongenerative grammatical models for document analysis," in *Computer Vision, 2005. ICCV 2005. Tenth IEEE International Conference on*, vol. 2, Oct 2005, pp. 962–969 Vol. 2.

[8] M. Rusinol, T. Benkhelfallah, and V. D'Andecy, "Field extraction from administrative documents by incremental structural templates," in *Document Analysis and Recognition (ICDAR), 2013 12th International Conference on*, Aug 2013, pp. 1100–1104.

[9] A. Lemaitre and J. Camillerapp, "HIP 2013 FamilySearch Competition - Contribution of IRISA," in *HIP - ICDAR Historical Image Processing Workshop*, Washington, United States, Aug. 2013.

[10] A. L. N. Fred and A. Jain, "Data clustering using evidence accumulation," in *Pattern Recognition, 2002. Proceedings. 16th International Conference on*, vol. 4, 2002, pp. 276–280 vol.4.

[11] C. Carton, A. Lemaitre, and B. Coüasnon, "Learnpos: a new tool for interactive learning positioning," in *Document Recognition and Retrieval DRR XXI*, 2014.

[12] A. O. Maroneze, B. Coüasnon, and A. Lemaitre, "Introduction of statistical information in a syntactic analyzer for document image recognition," in *DRR*, 2011, pp. 1–10.

[13] M. Garris, "Evaluating spatial correspondence of zones in document recognition systems," in *Image Processing, 1995. Proceedings., International Conference on*, vol. 3, Oct 1995, pp. 304–307 vol.3.